Our method is based on Lasso model and it includes Data processing, feature selection, model design, model selection, model ensemble.

We utilize RobustScaler to scale the data, add noise to prevent overfitting; utilize probing to select features; utilize Lasso to build model; utilize GridSearch to select models; ensemble selected using simple sampling. We obtained LB of 0.871 finally.

**Data processing:**

We use RobustScaler to scale attributes of both training data and test data. We find the data is close to Gaussian Distribution, but some of the data is deviated, so we use this method to scale data that are too deviated to avoid overfitting.

Then we add noise to training data attributes to avoid overfitting. The noise is subject to Gaussian distribution and is generated randomly.

**Feature selection:**

We utilize probing to do feature selection. Probing includes the follows steps:

1. Use simple model(like logistic regression) to make prediction with only 1 feature. Get 300 prediction file(each for one feature).
2. Submit there files to kaggle and collect scores.
3. Select features: if the score is higher than 0.5, select; else, abandon this feature.

We use person correlation and REFCV to select feature at the first step. And I realized that basically, feature selection is choose features relevant to the label and the submission score can reflect this relevance. So I come up with this method. Submission score higher than 0.5 means that the feature is relevant to the label.

**Model Design:**

Our method is based on Lasso model.

Lasso model add L1 penalty to standard linear model. Using the L1 penalty, the model can reduce some parameters to 0, which reduce the number of input dimension. So this model is suitable for this dataset.

**Model selection:**

Although the data processing method and model have been chosen, we still need to tune the hyper-parameters of this model to get best performance.

We choose Grid Search and Cross Validation to tune the model. In Cross Validation, we set k=20, which means the data is split into 20 sets, 19 are used to train, 1 is used to select model. We use regression score function to select model, the threshold is set to 0.190.

**Model Ensemble:**

we use the method of simple averaging to ensemble selected models.

We use selected models to make prediction and calculate the average as the final prediction.

Finally, our method get the public score of 0.871.